DOCUMENT RESUME

ED 146 229 TM 006 632

AUTHOR Rudnitsky, Alan N.
TITLE The Graphic Representation of Structure in Similarity/Dissimilarity Matrices: Alternative Methods.
PUB DATE [Apr 77]
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (61st, New York, New York, April 4-8, 1977)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Botany; *Cluster Analysis; Cluster Grouping; *Comparative Analysis; *Graphs; *Mathematical Models; *Matrices; *Multidimensional Scaling; Scientific Concepts; Validity
IDENTIFIERS *Hierarchical Clustering Analysis

ABSTRACT

Three approaches to the graphic representation of similarity and dissimilarity matrices are compared and contrasted. Specifically, Kruskal's multidimensional scaling, Johnson's hierarchical clustering, and Waern's graphing techniques are employed to depict, in two dimensions, data representing the structure of a set of botanical concepts. Each of the three representational techniques emphasizes a different aspect of structure in matrix data; specifically, content structure and cognitive structure are discussed. The researcher's assumptions regarding the structure of the data govern the choice of an appropriate representational method. When a hierarchical, class-inclusion structure is suspected, hierarchical clustering is the technique of choice. When underlying dimensionality is suspected or sought, and particularly when more than two dimensions are possible, multidimensional scaling is the proper technique. However, if there are no reasons for assuming the presence of a particular structure, Waern's graphing is the easiest and least expensive first choice. If a graphing solution yields promising results and two dimensions give adequate insight into structure, then combining the graph with a multidimensional scaling solution results in a representation with the greatest number of salient dimensions. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED146229

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Alan N. Rudnitsky

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

THE GRAPHIC REPRESENTATION OF STRUCTURE IN SIMILARITY/DISSIMILARITY MATRICES: ALTERNATIVE METHODS

Alan N. Rudnitsky
Smith College

TM006 632

This paper has been prepared for presentation at the annual meeting of the American Educational Research Association, New York, N.Y. 1977

The intent of this study is to compare and contrast three approaches to the graphic representation of similarity or dissimilarity matrices. Specifically, Kruskal's multidimensional scaling, Johnson's hierarchical clustering, and Waern's graphing techniques are employed to depict, in two dimensions, data representing the structure of a set of biological concepts.

The use of structure as a significant variable in educational research is becoming more and more widespread. Content, prose, or subject matter structure and cognitive structure are two such structural variables. (Johnson, 1967, 1969; Rothkoff, 1972; Geeslin, 1973; Shavelson, 1972, 1974; Rudnitsky, 1976; and Rudnitsky and Posner, 1976.) Typically the analysis of content and cognitive structure result in distance or similarity matrices containing data on the interpoint relationships of the items comprising the variable's structure. For the purposes of comparison, discussion and visualizability, these matrices are pictorially or graphically represented. This graphic representation is usually accomplished through the use of a multidimensional scaling or hierarchical clustering technique.

Waern (1972) has proposed a graphic method for representing structure. She notes that interpretation of dimensional methods (e.g., hierarchical clustering, multidimensional scaling) may be done with regard to configuration and underlying dimensionality. However, as Waern notes, the data often does not fulfill dimensional requirements and interpretation is limited to configuration.

The proposed graphing method aims at an output indicating the stimuli and their relations. This graph would be restricted to configurational interpretation. Waern describes the method and compares configurations obtained using it to those obtained using dimensional methods. Dimensional Method representations are found to be similar to Waern's graphing method with the latter being more easily interpretable.

Holman (1972) compared Johnson's (1967) hierarchical model for displaying psychological distances with Shepard's (1962) euclidean model. Holman finds these models to be theoretically mutually exclusive and exhaustive. He notes that

> ...the theoretical relationship between the two models is clear: they are in fact at opposite extremes. Specifically, a set of errorless distance data will never satisfy both models nontrivially, but will always satisfy one of the models to some extent at least. Of course, real data will usually contain errors and thus will not fit either model exactly.[1]

Rapoport and Fillenbaum (1972) in looking at the applicability of Kruskal's (1964) multidimensional scaling and Johnson's (1967) hierarchical clustering find scaling to be more appropriate for linear underlying structures while clustering is best employed to represent taxonomic structures.

Preece (1976) compared various methods of measuring cognitive structure. Citing the difficulty of deciding, a priori, whether physics concepts (i.e., mechanics) could best be represented spatially (i.e., multidimensional scaling) or hierarchically (i.e., hierarchical clustering), Preece employed Waern's graphing method. His study compares various measures of a cognitive structure rather than methods used to visually depict the data. Waern's method did, however, depict the concepts meaningfully.

---

[1]Holman, E.W. "The Relation Between Hierarchical and Euclidean Models for Psychological Distances." Psychometrika, Vol. 37, No. 4, 1972, p. 419

The representational techniques (i.e., Waern's graphing, hierarchical clustering, multidimensional scaling) have never been compared in their representation of a structure other than that of physics or mathematics concepts, both of which have rather clearly defined relationships among the concepts.

This study compares structural representations obtained using Kruskal's multidimensional scaling (Version V), Johnson's hierarchical clustering, and Waern's graphing methods. The study includes data on both content and cognitive structures which allows for limited theoretical as well as purely visual comparison. Botany concepts are used, making prior assumptions about structure or underlying dimensionality difficult.

METHOD

In previous studies, Rudnitsky (1976) and Rudnitsky and Posner (1976), the relationship between content and cognitive structure was examined. A two-year college botany unit, plant growth and development, was sequenced in two ways. Content structure was then assessed for each sequence. The cognitive structure of students using the respective units was also assessed. Content and cognitive structure were each assessed in several ways. The most meaningful results, those that discriminated between the cognitive structure of groups and prose structure of the two sequences in the most interpretable manner, were obtained by having judges "graph" content structure and by assessing cognitive structure through a graph construction test. These measures yielded matrices of interpoint distances among the seventeen botany concepts. The matrices were visually represented as 17 points in two dimensional

space using Kruskal's multidimensional scaling. The study focused on the validity of cognitive structure measures and the relationship of content and cognitive structure.

In order to compare hierarchical clustering and graphing to multidimensional scaling the Rudnitsky (1976) data is represented by each technique. As well as being able to compare identical matrices represented by the three techniques, cognitive and content structure can be compared within each technique to further explore the technique's usefulness.

The multidimensional scaling solutions were obtained using Version V of Kruskal's MDSCAL (Rao, 1972). Stress values range from 0.049 to 0.292 for the scaling solutions. Kruskal (1968) suggests the following verbal evaluation of goodness-of-fit of a configuration to the data:

Goodness of Fit	Stress
Poor	.40
Fair	.20
Good	.10
Excellent	.05
"Perfect"	.00

However, Kruskal notes that:

> ...under some circumstances this table is not at all applicable and, at best, it is only a rough guide. For example, where data values are heavily replicated, this table is pessimistic and larger stress values are acceptable. For some situations there is simply no experience on which to base any evaluation.[2]

Higher stress, two-dimensional solutions were used rather than lower stress three and four-dimensional solutions for several

---

[2] Kruskal, J.B. "How to Use M-D Scale, a Program to do Multidimensional Scaling and Multidimensional Unfolding," in Computer Programs for Multidimensional Scaling, V.R. Rao, Cornell University, Ithaca, New York, 1972.

reasons. First, many of the matrice's contain heavily replicated values. Second, two-dimensional solutions are more easily visualizable and interpretable than are higher dimensional solutions and are more easily compared to the other representations. Since for botany concepts the identification of configurations and clusters is a more realistic task than discovering underlying dimensions, any solution accurately portraying these clusters is acceptable. As a check, all solutions of stress higher than .20 were also plotted in three-dimensions and collapsed to two dimensions. The clusters obtained did not change.

This study employs Johnson's (1967) Hierarchical Clustering Scheme using the diameter or maximum method of computation.

The graphing technique used is the multi-step method described by Waern (1972). Cut-offs are indicated on each graph.

The seventeen botany concepts and their abbreviations are listed in Table 1.

Table 1 Abbreviations of Concepts

SD - seed	GM - germinate
RT - root	M/D - Monocot/dicot
SM - stem	MER - meristematic region
STR - structure	ELN - elongation region
TS - tissue	MAT - maturation region
FN - function	CLD - cell division
GW - grow	CLE - cell elongation
CL - cell	CLM - cell maturation
SPC - specialization	

Results

Any method of depicting the structure of similarity/dissimilarity data results in a representation which is suggestive rather than ironclad or "true." The issue, then, is the insight into structure afforded a researcher by employing a particular technique with particular data. That is to say, what matters to the researcher is what the representation looks like.

The results are presented here so that initial comparisons made are those between methods for a particular set of data. Verbal description is kept to a minimum. Some introductory remarks are made to assist the reader in interpreting the representations.

Content Structure

The content structure depicted here is that of a World-related script and that of a Concept-related script. The World-related script organized the seventeen concepts on the basis of spatial and temporal plant characteristics. This version first discussed the seed, then the root, and lastly the stem of vascular plants; these being spatial characteristics. Temporally a plant develops from a seed, the root is the first emergent plant part followed by the stem. Structure, tissues, function, and growth are discussed for each plant part. The Concept-related script organizes the unit around "higher order" concepts (identified by subject matter specialists). This version deals with plant structures, tissues, growth, and functions. The appropriate aspect of plant parts (i.e., seeds, roots, and stems) are discussed together for each major concept.

Key aspects in these representations are the growth regions (i.e., meristematic, elongation, and maturation regions), the growth

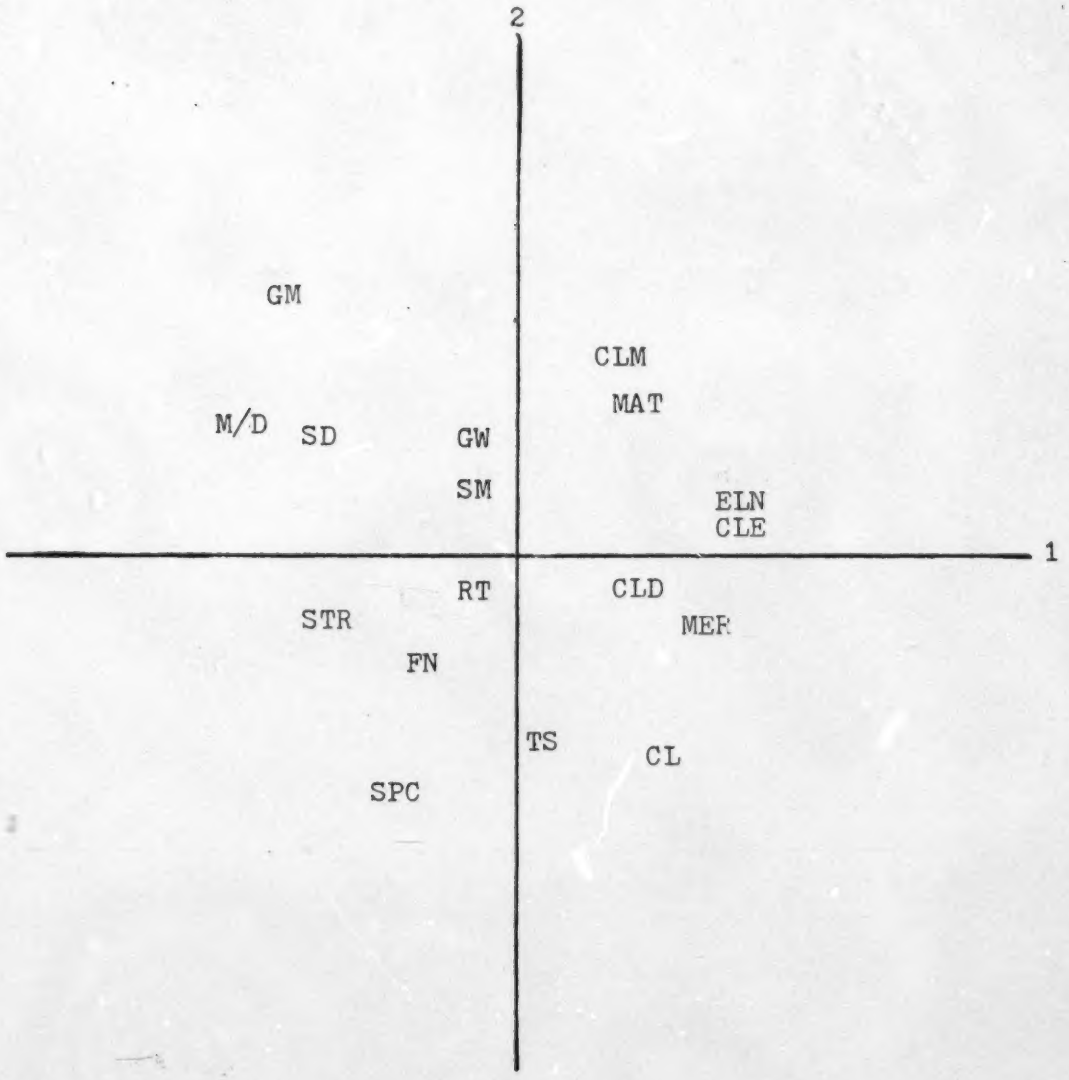
processes (i.e., cell division, cell elongation, cell maturation), and the positions of seed, root, stem in relation to growth, structures, tissues, and functions.

Cognitive Structure

The cognitive structure representations that follow are those assessed using a graph construction test. The groups (i.e., World-related, Concept-related) correspond to the script used as the instructional treatment.

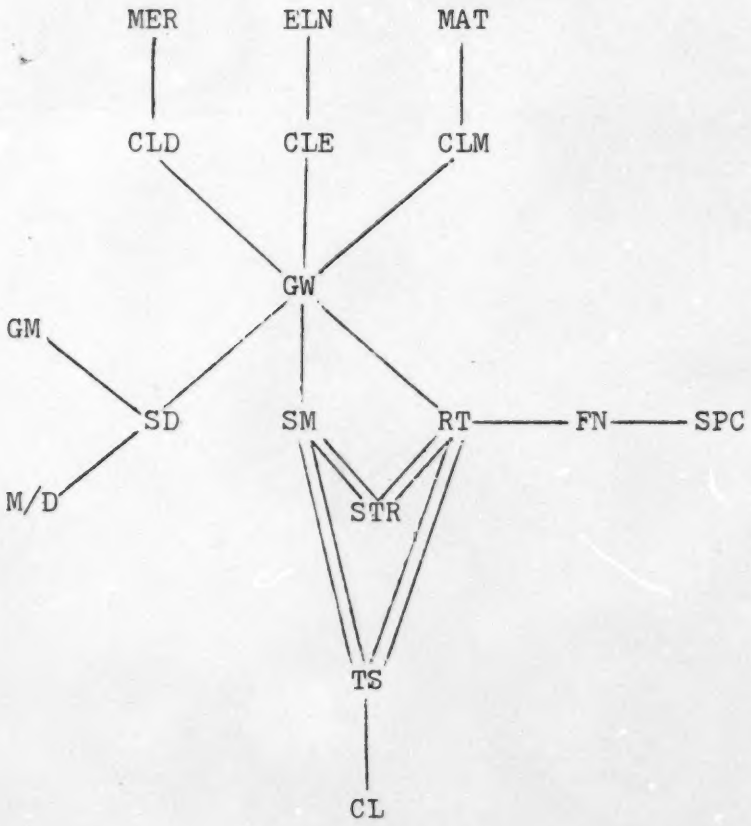
Multidimensional Scaling Solution--Content Structure
for World-Related Script

stress = .187

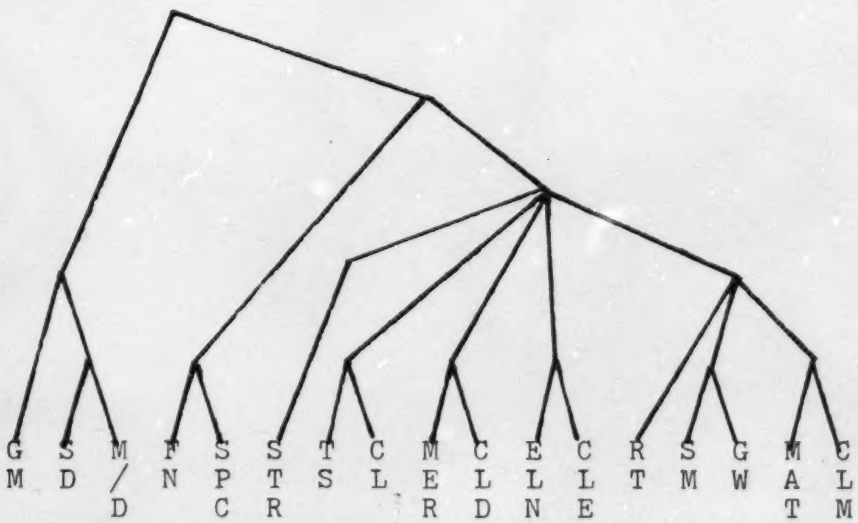


Waern's Graphing Approach--Content Structure for World-Related Script

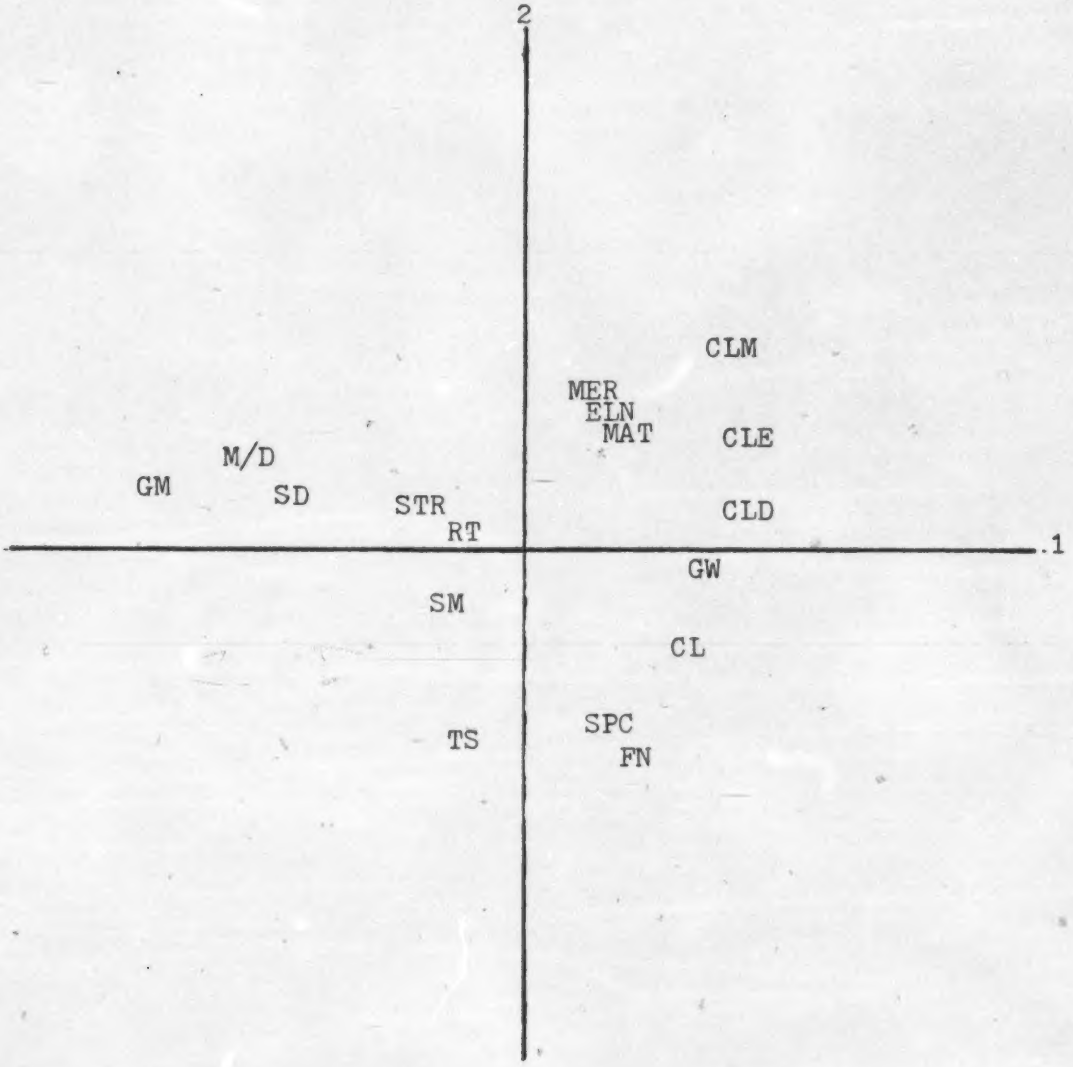
cut-offs:
1.5 ——
2.5 ═══



Hierarchical Clustering Solution--Content Structure
for World-Related Script

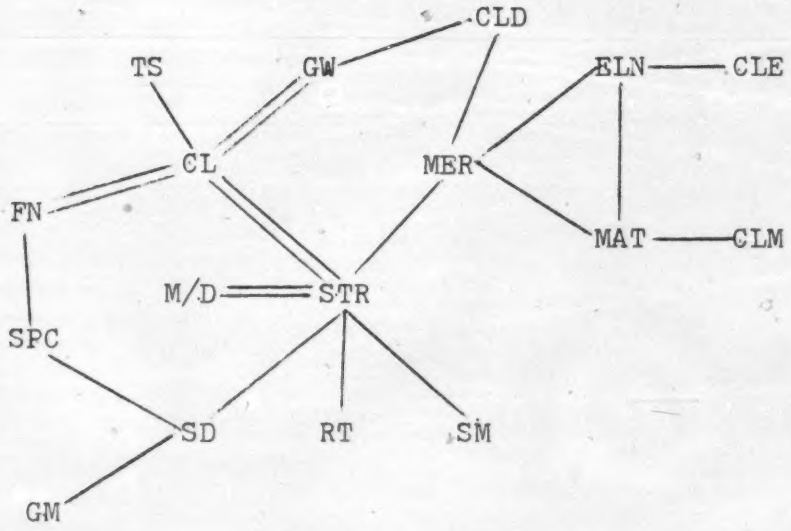


Multidimensional Scaling Solution--Content Structure
for Concept-Related Script

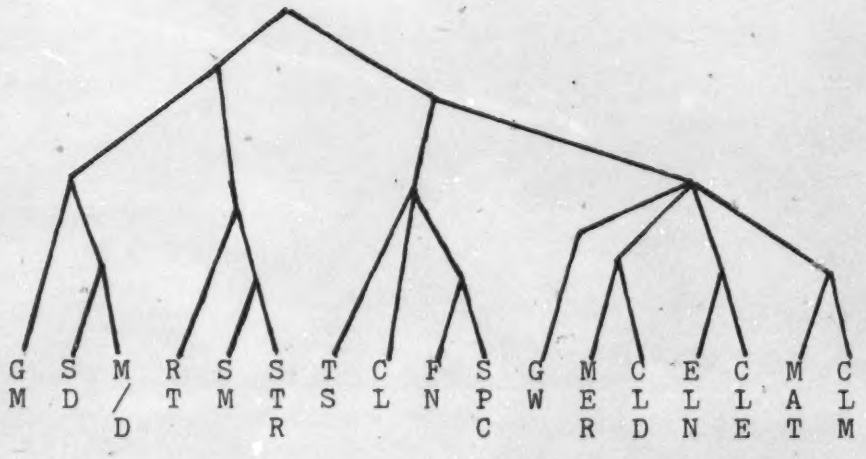


Waren's Graphing Approach--Content Structure for Concept-Related Script

cut-offs:
1.5 ——
2.5 ═══

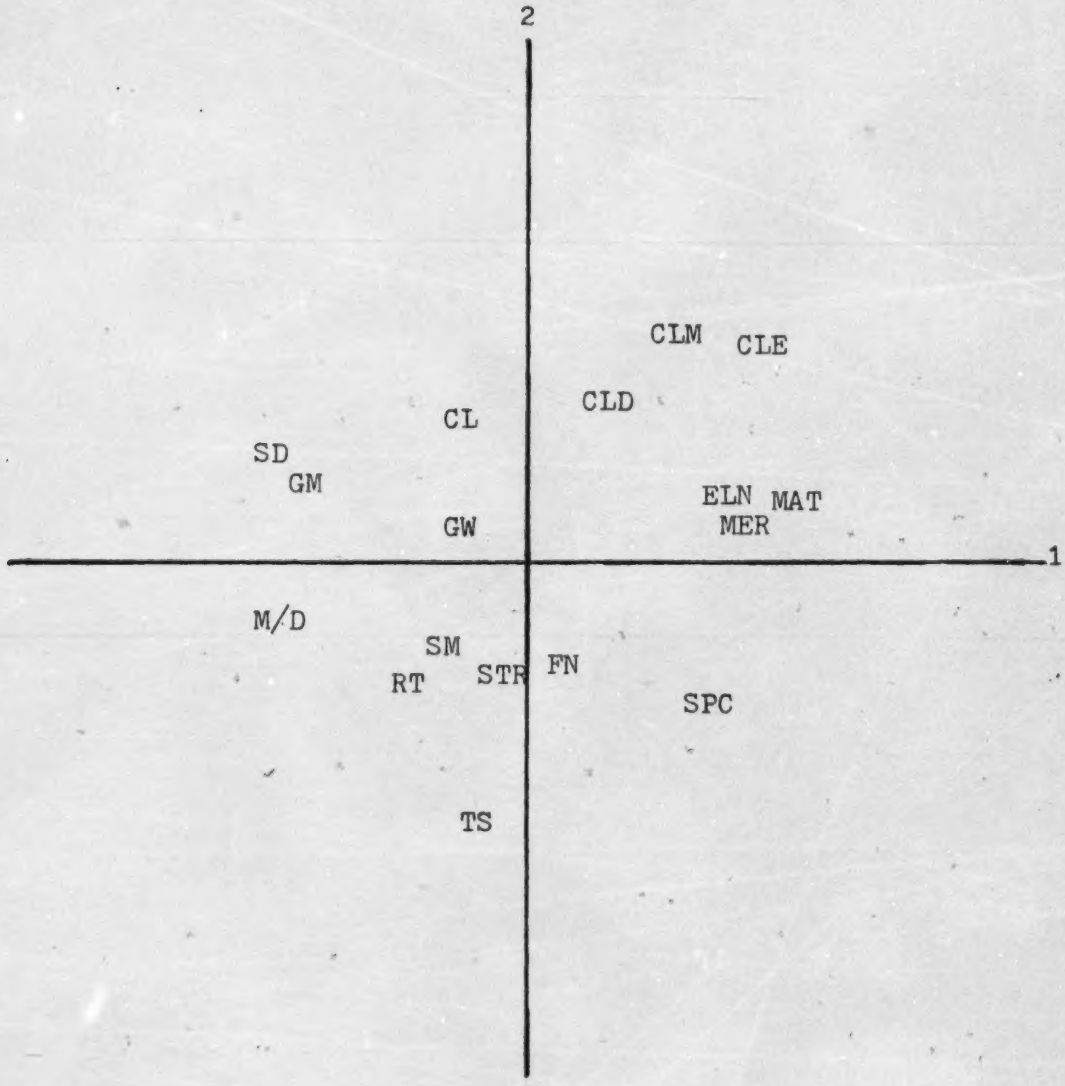


Hierarchical Clustering Solution--Content Structure
for Concept-Related Script



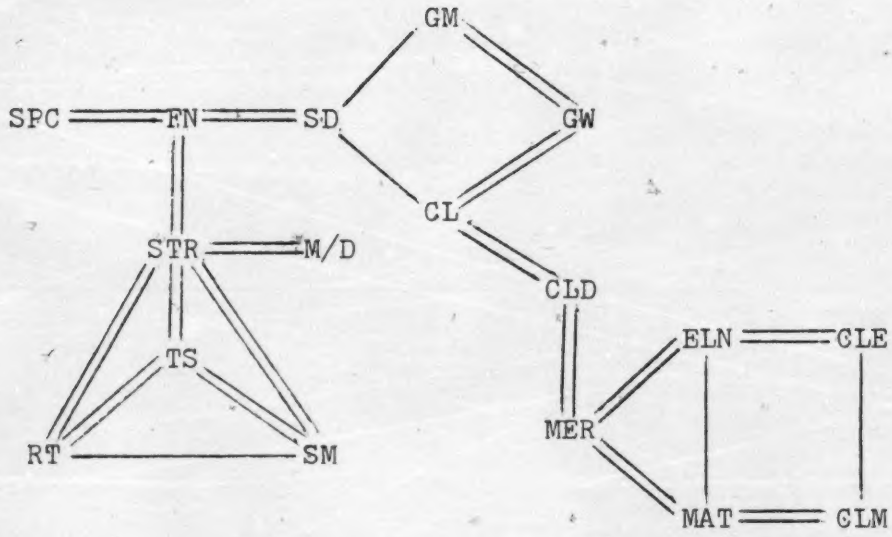
Multidimensional Scaling Solution--Cognitive Structure
for World-Related Group

stress = .253

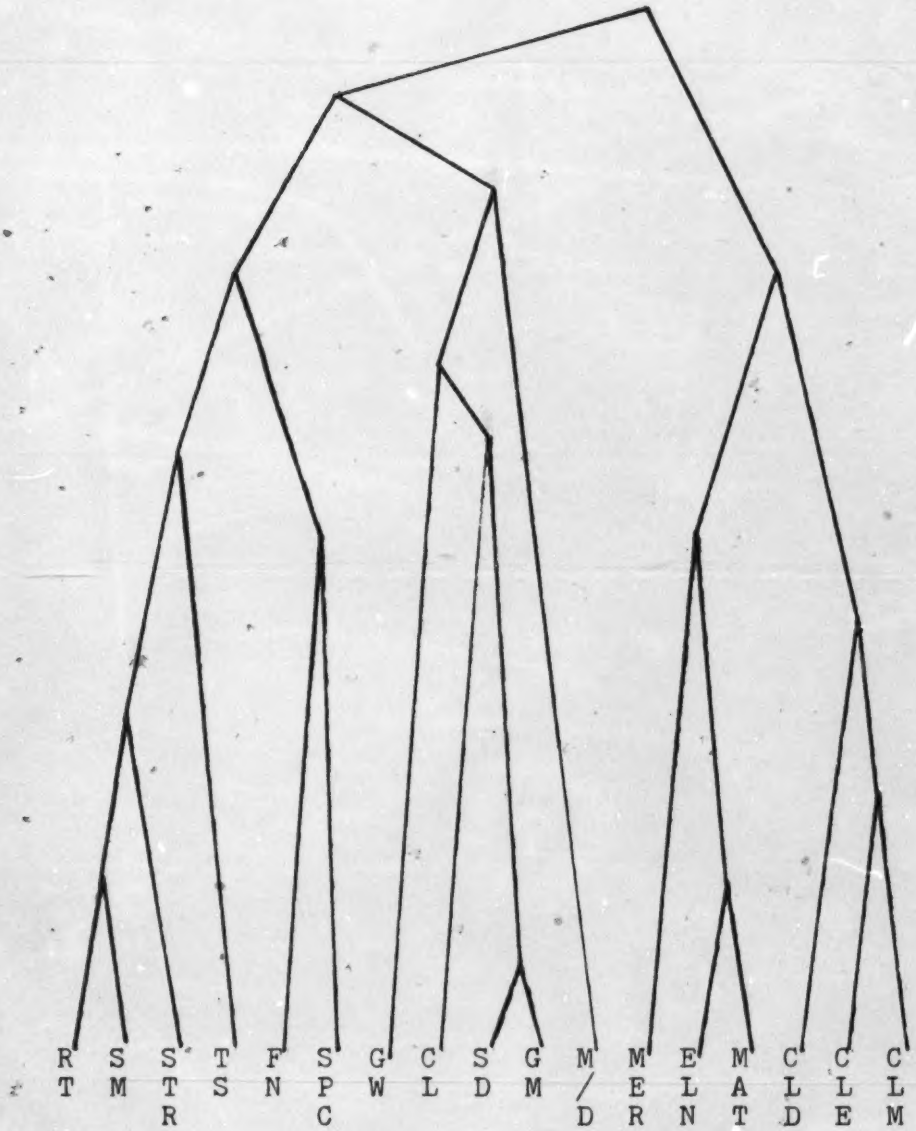


Waern's Graphing Approach--Cognitive Structure for World-Related Group

cut-offs:
2 ——
3 ═══

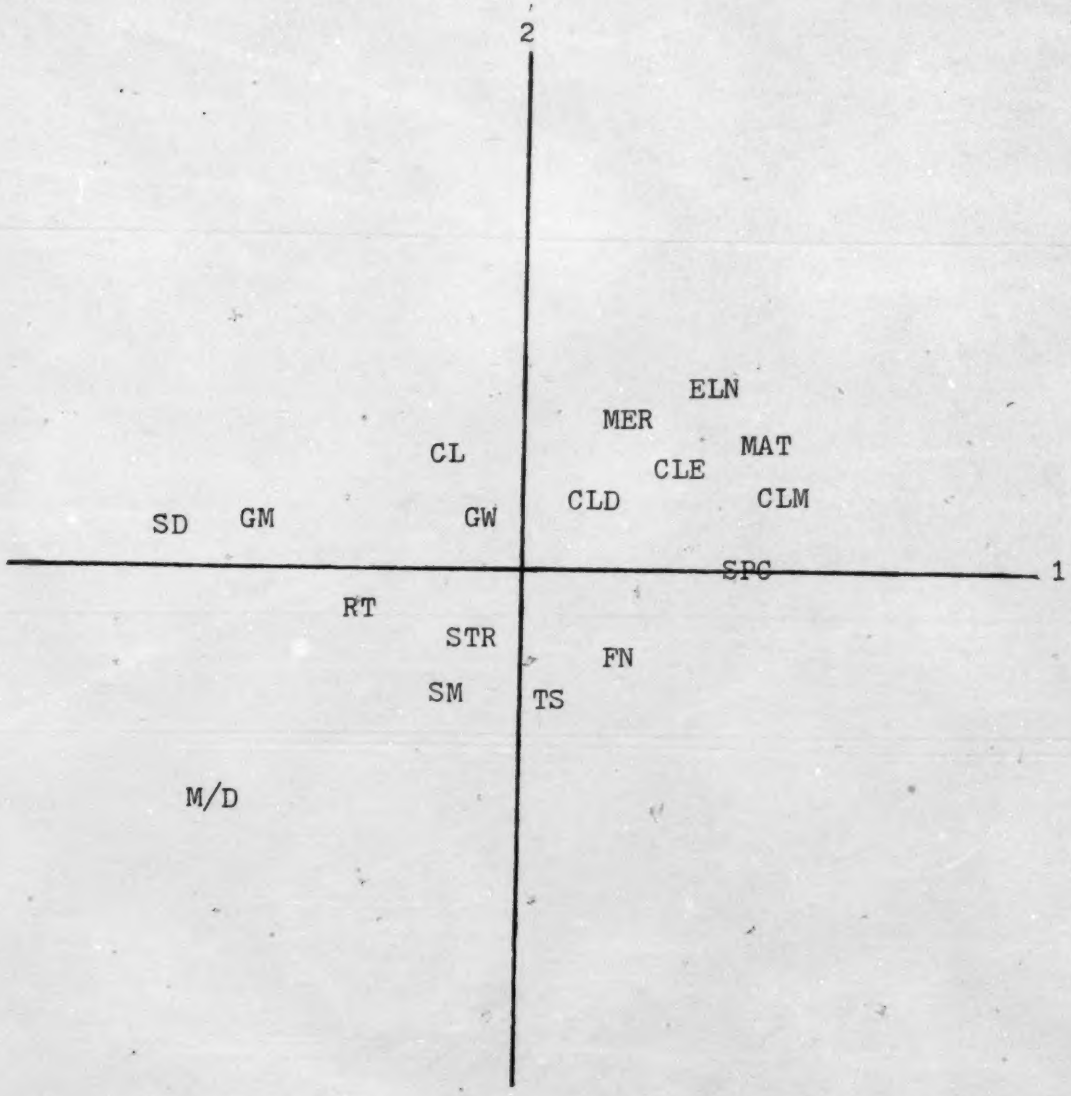


Hierarchical Clustering Solution--Cognitive Structure
for World-Related Group

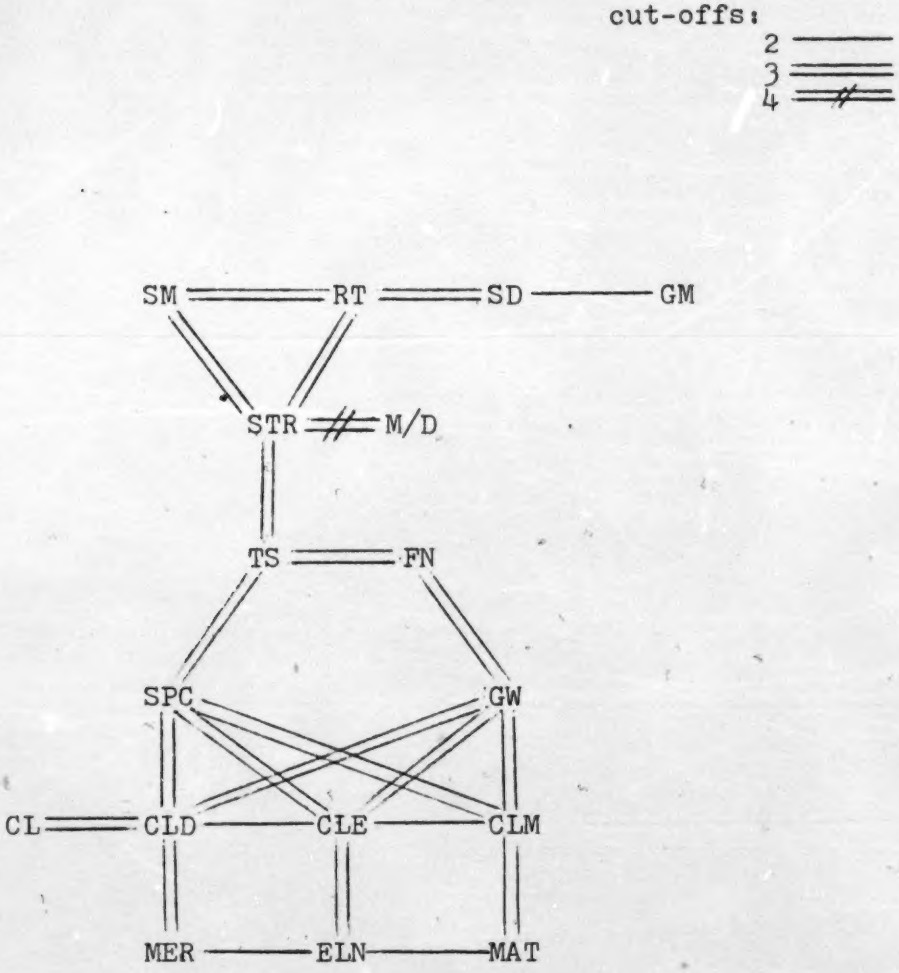


Multidimensional Scaling Solution--Cognitive Structure
for Concept-Related Group

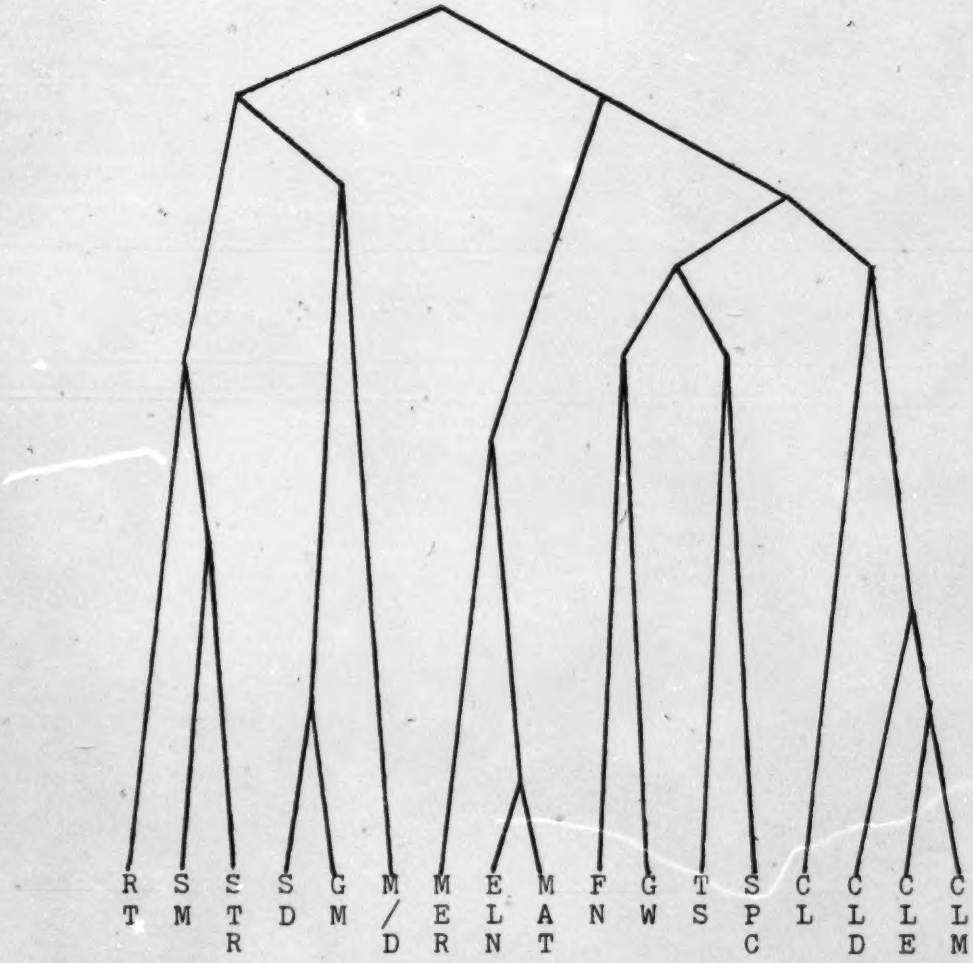
stress = .292



Waern's Graphing Approach--Cognitive Structure for Concept-Related Group



Hierarchical Clustering Solution--Cognitive Structure
for Concept-Related Group



Interpretation

Interpretation of graphic representations are based on the insights they afford into structure. To help determine this, representations are compared within a particular data set (i.e., matrix) and representations are compared across data sets.

The salient structural feature depicted by each method varies. Hierarchical clustering emphasizes class inclusion, that is taxonomic or categorical features of structure. As well as what clusters or categories are formed, hierarchical clustering gives some insight into "strength" of clusters, how "quickly" concepts cluster together.

Multidimensional scaling emphasizes clusters and, to a lesser extent, dimensionality. Clusters of concepts can be interpreted in terms of their similarity or their membership in a class. Another feature of the clustering is the interrelationships among clusters, that is, their relative distances. Where there is a good fitting (low stress) solution in N-dimensions it is reasonable to try to determine what the underlying dimensions of structure might be.

Waern's graphing approach emphasizes relationships among concepts. Relative distance among concepts and, to some extent, their orientation are at the discretion of the researcher. The lines which form connections are the important aspect of the graphing approach.

Visual examination of the three representations for each data set reveals several similarities. Multidimensional clusters, related concepts in the graphing approach, and classes formed in hierarchical clustering are, to a great extent, similar. That is, if one wanted to form groups of two or three concepts on the basis of proximity or connectedness, these groups would be similar for each technique.

In terms of overall interpretability, hierarchical clustering for this data is less useful than either multidimensional scaling or Waern's graphing. The botany concepts do not have a taxonomic structure (or at least one which is readily apparent and applicable to all solutions), thus insight into overall content or cognitive structure is not greatly facilitated by hierarchical clustering.

Graphing and multidimensional scaling do suggest an overall structure. Scaling indicates relative position or proximity and to some extent indicates which elements are central in structure and which are peripheral. Graphing suggests the degree of connectedness and centrality of the elements in the structure.

Another desirable feature of scaling and graphing is the ability to compare representations of, for example, content and cognitive structure. A researcher desiring insight into the relationship of content and cognitive structure in terms of which concepts are central or in similar clusters or possessed of similar relationships could use scaling or graphing.

Both multidimensional scaling and Waern's graphing have limitations. On the one hand, multidimensional scaling lacks lines that indicate relationships or linkages, a strength of Waern's approach. Waern's graphing, however, leaves orientation of the configuration very much at the discretion of the researcher. Related to this is the fact that relative proximity has no (or little) significance in the graphing approach, while this is not the case for multidimensional scaling.

A combination of the above two techniques in which the strengths of each is shared with the other has been explored. Two variations

of this combination are possible.[3] One approach would be to use the relationships indicated in a graphing solution to impose lines on a multidimensional scaling solution (see Figure A). A second approach would be to orient and space a graphing solution using the information available in a multidimensional scaling solution. This process involves placing concepts on the graph in varying proximity and making orientation decisions based on scaling solutions (Figure B). Both approaches result in very similar, in fact, virtually identical representations. While not geometrically neat, these representations do combine the strengths of scaling and graphing and result in a "picture" that is capable of being compared to other "pictures" in terms of relationships, clusters, proximity, and orientation.

---

[3]The author wishes to thank George Posner and Leon Pines for their helpful suggestions in this and other matters. All shortcomings of the paper are solely this author's responsibility.

Figure A. Multidimensional Scaling Solution with added lines--
Cognitive Structure for World-Related Group

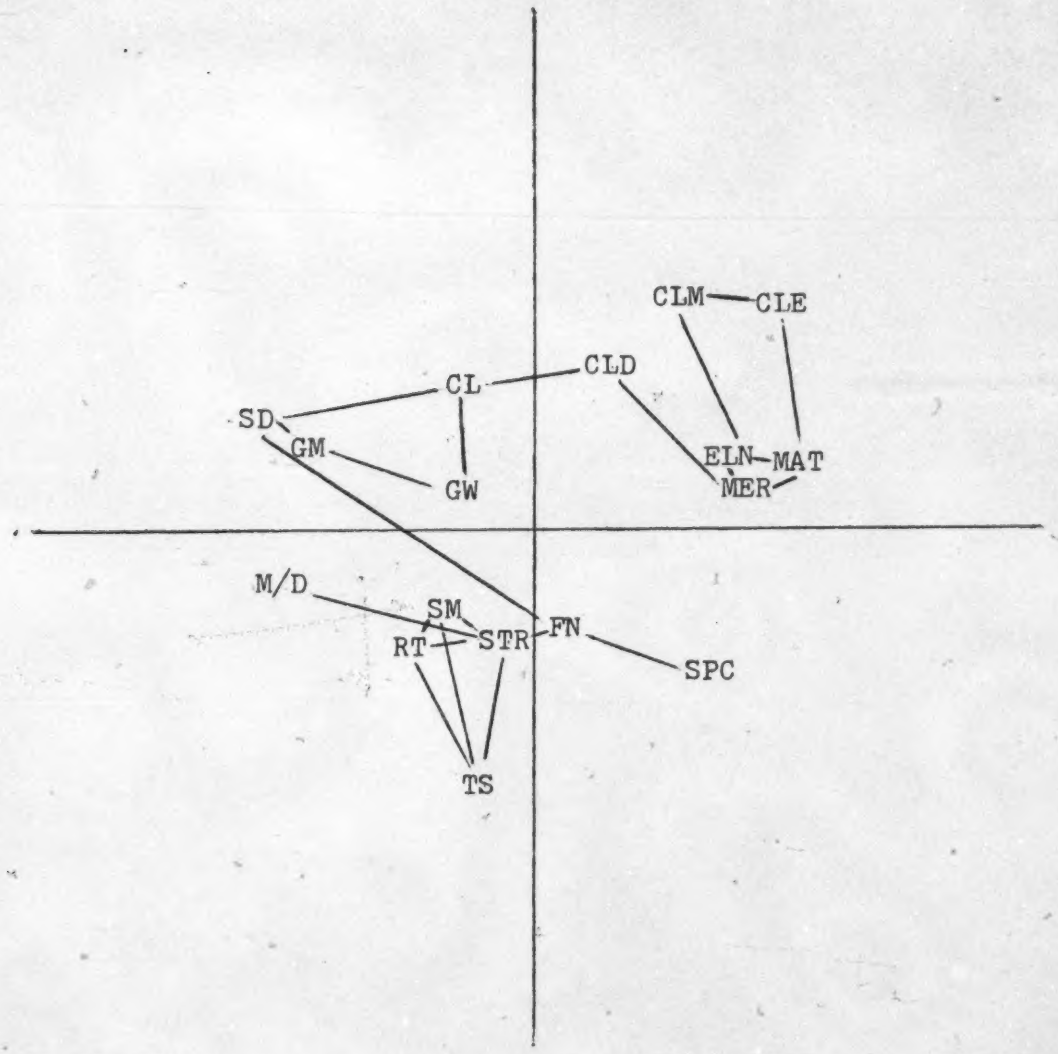
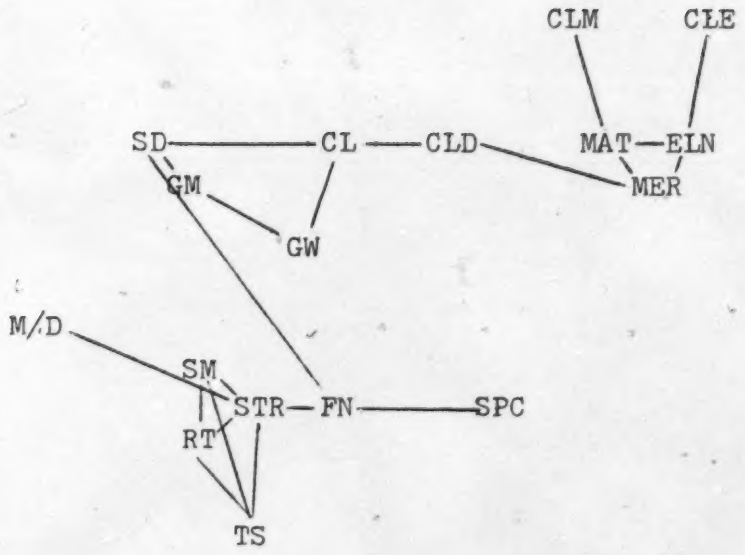


Figure B. Waern's Graphing Solution oriented and spaced using information from multidimensional scaling solution--Cognitive Structure for World-Related Group.



Conclusions

Each of the three representational techniques emphasize a different aspect of structure in matrix data. Visual comparison of the representations in this paper do speak more eloquently than any verbal description which could be appended. For this researcher, Waern's graphing combined with scaling gives the best and most fruitfully compared insights into structure.

In terms of choosing an appropriate representational method, alot rests on the researcher's assumptions regarding the structure of the data. If a hierarchical, class-inclusion structure is suspected, one in which items group together in more inclusive classes, hierarchical clustering is the technique of choice. If underlying dimensionality is suspected or sought, particularly if more than two dimensions are possible, multidimensional scaling is the technique of choice. On the other hand, if one has no a priori reasons for suspecting a particular structure, Waern's graphing is the easiest and least expensive first choice.

If a graphing solution yields good or promising results and two dimensions give adequate insight into structure, then combining the graph with a multidimensional scaling solution results in a representation with the greatest number of salient dimensions.

Further comparisons of these techniques, particularly if done with content representing a wide cross-section of disciplines, would be a useful next step and would provide more guidance to a researcher in the position of having to choose a technique to graphically depict similarity/dissimilarity matrix data.

Bibliography

Geeslin, W.E. "An Exploratory Analysis of Content Structure and Cognitive Structure in the Context of a Mathematics Instructional" (Doctoral Dissertation, Stanford University), Ann Arbor, Michigan: University Microfilms, 1973 (74-6478).

Holman, E.W. "The Relation Between Hierarchical and Euclidean Models for Psychological Distances," Psychometrika, December 1972, Vol. 37, No. 4, 417-423.

Johnson, S.C. "Hierarchical Clustering Schemes," Psychometrika, September 1967, Vol. 32, No. 3, 241-254.

Kruskal, J.B. "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," Psychometrika, March 1964, Vol. 29, No. 1, 1-27.

Preece, P.F.W. "Mapping Cognitive Structure: A Comparison of Methods," Journal of Educational Psychology, 1976, Vol. 68, No. 1, 1-8.

Rao, V.R. Computer Programs for Multidimensional Scaling, Ithaca: Cornell University, 1972, 1-43.

Rapoport, A. & Fillenbaum, S. "An Experimental Study of Semantic Structures," in Romney, A.K., Shepard, R.N. and Nerlove, S.B. (eds.) Multidimensional Scaling: Theory and Applications in the Behavioral Sciences (Vol. 2), New York: Seminar Press, 1972, 96-131.

Rothkopf, E.Z. "Structural Text Features and the Control of Processes in Learning from Written Materials," in Freedle, R.O. & Carroll, J.B. (eds.) Language Comprehension and the Acquisition of Knowledge, Washington D.C.: V.H. Winston and Sons, 1972, 315-334.

Rudnitsky, A.N. Content Structure, Cognitive Structure and Their Relationship: A Methodological Investigation. Unpublished doctoral dissertation, Cornell University, 1976.

Rudnitsky, A.N. & Posner, G.J. "The Relationship of Content Structure and Cognitive Structure," paper presented at AERA annual meeting, San Francisco, 1976.

Shavelson, R.J. "Some Aspects of the Relationship Between Content Structure and Cognitive Structure in Physics Instruction," (doctoral dissertation, Stanford University) Ann Arbor, Michigan: University Microfilms, 1971 (71-19, 259).

Shepard, R.N. "Analysis of Proximaties: Multidimensional Scaling with an Unknown Distance Function," I. Psychometrika, 1962a, 27, 125-140.

Shepard, R.N. "Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function," II. Psychometrika, 1962b, 27, 219-246.

Waern, Y. "Structure in Similarity Matrices: A Graphic Approach," Scand. J. Psychol., 1972, Vol. 13, 5-16.